MULTIMEDIA UNIVERSITY

STUDENT ID

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 3, 2015/2016

TPA7021 – Data Preprocessing and Analysis

(All sections / Groups)

23 MAY 2016
8.00 p.m - 10.00 p.m
(2 Hours)

INSTRUCTIONS TO STUDENTS

1. This question paper consists of 6 pages, including the cover page, with FOUR questions only.

2. Attempt **ALL** questions. All questions carry equal marks and the distribution of the marks for each question is given.

3. Please write all your answers in the answer booklet provided.

Answer ALL Questions

QUESTION 1

(a) Real world data is often "dirty". Discuss THREE (3) characteristics and sources of dirty data.

[3 marks]

(b) Name and explain briefly the phases involved in data preprocessing.

[3 marks]

(c) Data checking is often needed in data preprocessing. The `ifelse()` function in `R` programming can act as a "flag" to data that has violated the constraints. Study the data frame `dt.CGPA` below carefully and write an R script to check which row has CGPA > 4.00. Set "1" for CGPA>4.00, otherwise "0". Create a new variable named `chkCGPA` to store the results.

Student_ID	Student_Name	CGPA
100	Ting	5.20
200	Chong	3.76
300	Fatimah	3.90

[2 marks]

(d) Define *outlier* in the context of data preprocessing. Discuss whether all outliers should be treated as dirty data.

[2 marks]

QUESTION 2

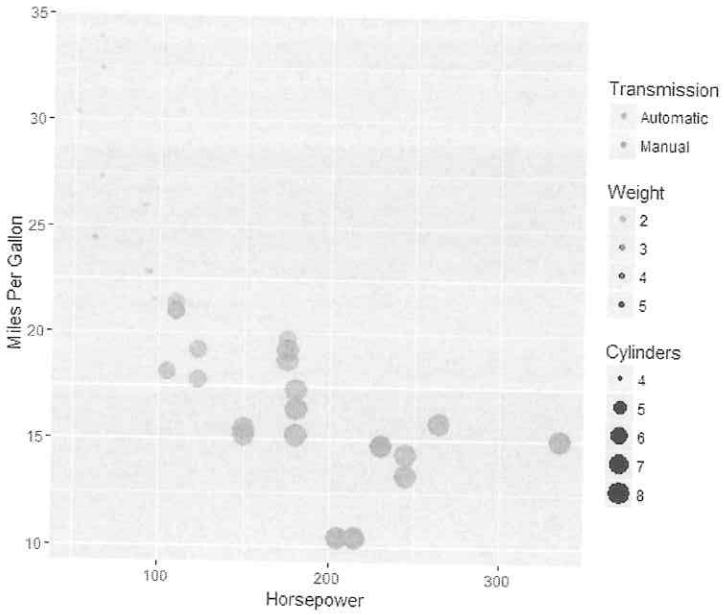
(a) Assuming that a data analyst would like to visualize the relationship between two variables X and Y. The only information given is that the variables are having numerical data. What plot/chart should the data analyst use? Give a reason for that.

[3 marks]

Continued...

(b) What chart should one use to describe the customers' behavior in navigating website? [2 marks]

(c) Study the following chart carefully. The dataset used is `mtcars`.

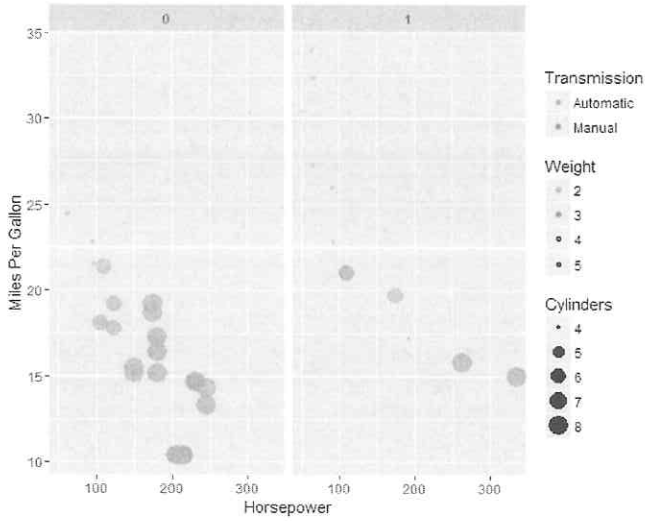


The R script to generate the above chart is as follows:

```
ggplot(data = mtcars, aes(x = hp, y = mpg,color =
factor(am), alpha = wt, size = cyl)) + geom_point() +
scale_color_discrete(labels = c("Automatic",
"Manual")) + labs(color = "Transmission",
x = "Horsepower", y = "Miles Per Gallon",
alpha = "Weight", size = "Cylinders")
```

Continued...

How should the R script above be extended to separate the chart into two separate charts based on *Transmission*? The resulting chart is shown below:



[1.5 marks]

(d) Study the following R codes carefully.

```
> Group1 <- c(2,3,7,2,6)
> Group2 <- c(10,8,7,5,10)
> Group3 <- c(10,13,14,13,15)
> Combined_Groups <-
  data.frame(cbind(Group1, Group2,Group3))
> Stacked_Groups <- stack(Combined_Groups)
> Anova_Results <- aov(values ~ ind,
                       data = Stacked_Groups)
> TukeyHSD(Anova_results)

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = values ~ ind, data = Stacked_Groups)

$ind
               diff       lwr       upr     p adj
Group2-Group1    4 0.4206853  7.579315 0.0286585
Group3-Group1    9 5.4206853 12.579315 0.0000598
Group3-Group2    5 1.4206853  8.579315 0.0075279
```

As a data scientist, what can you conclude?

[2 marks]

Continued...

(e) Name three methods for imputation of missing data during data preprocessing.

[1.5 marks]

QUESTION 3

(a) Briefly describe each of the heterogeneity problems with an example:
   i. Schema heterogeneity
   ii. Data type heterogeneity
   iii. Value heterogeneity

[6 marks]

(b) A trainer wants to determine the relationship between the years of working experience (X) and the expected score in a test (Y). The following shows the results of the linear regression model in R, with Y as the outcome and X as the predictor:

```
Call:
lm(formula = Y ~ X)

Residuals:
   Min     1Q Median     3Q    Max
-7.248 -2.902 -1.652  2.919 10.752

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   31.533      6.360   4.958 0.001110 **
X             10.905      1.744   6.253 0.000245 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.651 on 8 degrees of freedom
Multiple R-squared:  0.8301, Adjusted R-squared:  0.8089
F-statistic: 39.09 on 1 and 8 DF,  p-value: 0.000245
```

   i. Based on the regression model above, find the expected test score if a student has 5 years of working experience. Show your working.

[2 marks]

Continued...

ii. What is the value of correlation `r`? Show your working.

[1 mark]

iii. What does the value in Q3(b)(ii) tell you about the linear relationship between the years of working experience and test score?

[1 mark]

QUESTION 4

(a) Heuristic search algorithms such as the greedy forward selection or greedy backward elimination are often employed for attribute subset selection. Why are these algorithms called greedy?

[2 marks]

(b) The following table shows the data set of a football match analysis:

Fouls	Possession	Shots	Result
No	Yes	Strong	Win
Many	Yes	Strong	Win
Few	No	Strong	Win
Many	Yes	Weak	Loss
Few	Yes	Weak	Unknown
Few	No	Strong	Win
Many	No	Weak	Draw
No	Yes	Weak	Draw
Many	Yes	Strong	Win
No	No	Weak	Draw

i. Given that `Result` is the class attribute, what is the entropy of the data set? Show your working.

[2 marks]

ii. Using the ID3, which attribute has the maximal information gain? Show your working.

[6 marks]

End of Page